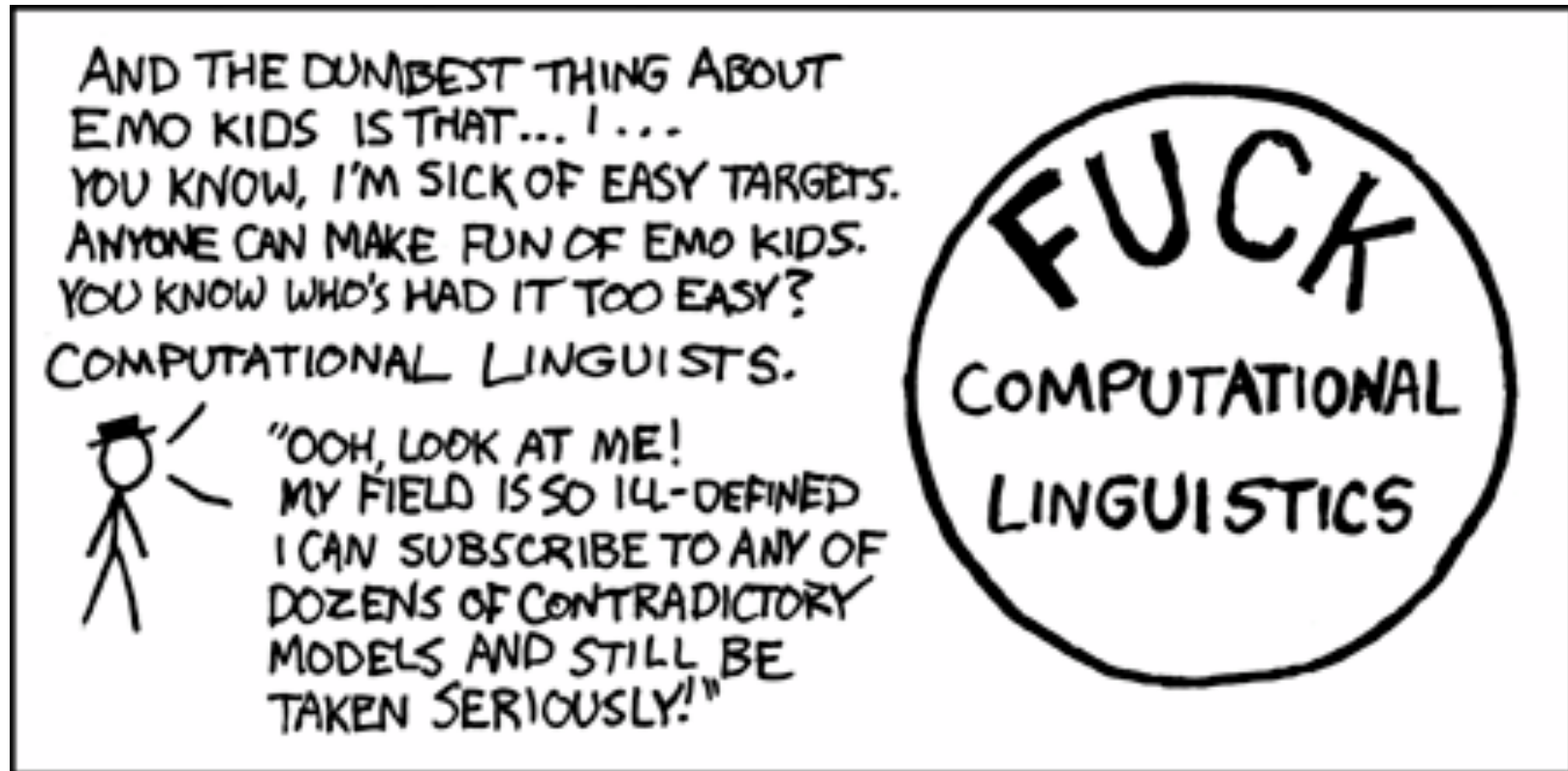


Computational Linguistics (aka Natural Language Processing)

Bill MacCartney
SymSys 100
Stanford University
26 May 2011

(some slides adapted from Chris Manning)

xkcd snarkiness



OK, Randall, it's funny ... but wrong!

A word on terminology

If you call it ...

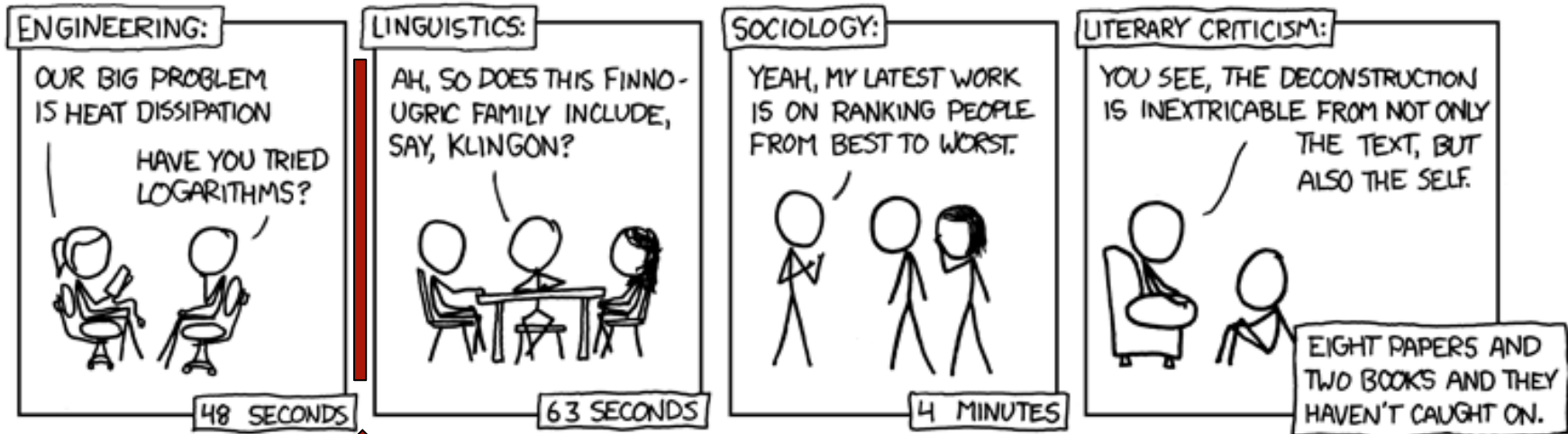
- **Computational Linguistics (CL)**
 - ... you're a linguist!
 - ... you use computers to study language
- **Natural Language Processing (NLP)**
 - ... you're a computer scientist!
 - ... you work on applications involving language

But really, they're pretty much synonymous

Let's get situated!

MY HOBBY:

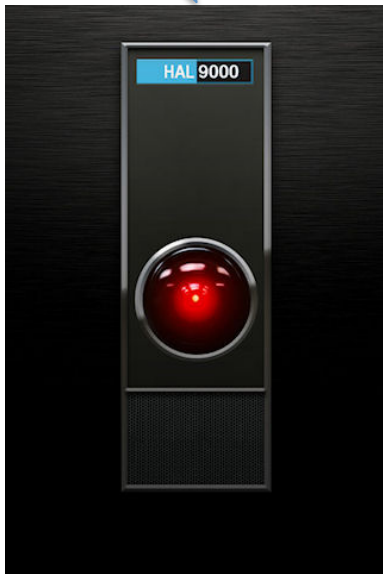
SITTING DOWN WITH GRAD STUDENTS AND TIMING HOW LONG IT TAKES THEM TO FIGURE OUT THAT I'M NOT ACTUALLY AN EXPERT IN THEIR FIELD.



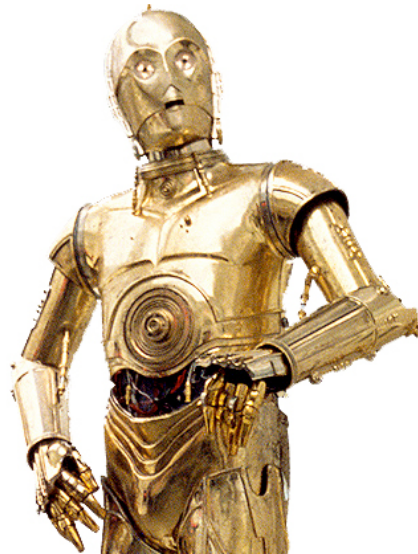
Today, we are in this interstice

NLP: The Vision

I'm sorry, Dave.
I can't do that.



Oh, dear!



That is correct,
captain.



Language: the ultimate UI



Where is **A Bug's Life** playing in **Mountain View**?

A Bug's Life is playing at the Century 16 Theater.

When is **it** playing **there**?

It's playing at 2pm, 5pm, and 8pm.

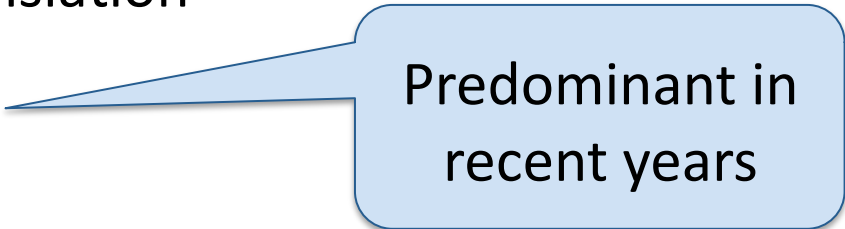
OK. I'd like 1 **adult** and 2 **children** for **the first show**.
How much would **that** cost?



But we need **domain knowledge**, **discourse knowledge**, **world knowledge**
(Not to mention linguistic knowledge!)

NLP: Goals of the field

- From the lofty ...
 - full-on natural language understanding
 - participation in spoken dialogues
 - open-domain question answering
 - real-time bi-directional translation
- ... to the mundane
 - identifying spam
 - categorizing news stories (& other docs)
 - finding & comparing product information on the web
 - assessing sentiment toward products, brands, stocks, ...



Predominant in recent years

NLP in the commercial world

The image features a collage of various logos and interface elements:

- Calendar Interface:** A snippet of a calendar showing a date range from 1:15-2:05 for a session titled "the problem session for your course, CS224n in Gates B03 from Skilling Aud". A dropdown menu is open with options: "Create New iCal Event..." and "Show This Date in iCal..."
- Logos:** cymfony (harnessing influ...), PEARSON Knowledge Technologies, Google (with "ski resort" in a search box), Powerset (a question mark in a circle), YAHOO!, Microsoft, hakia (BETA search for meaning), collective intellect, J.D. POWER AND ASSOCIATES®, and Nielsen BuzzMetrics.
- Search Results:** A snippet of search results for "ski resort" showing "Sugar Bowl Ski Lodging" with a description: "Escape to our Snowbound Village. Fresh tracks steps from your room." and the URL "www.sugarbowl.com".
- Text Snippet:** A box containing the text "Their are many approaches, out" with a green wavy underline under "are".

Current motivations for NLP

What's driving NLP? Three trends:

- The explosion of machine-readable natural language text
 - Exabytes (10^{18} bytes) of text, doubling every year or two
 - Web pages, emails, IMs, SMSs, tweets, docs, PDFs, ...
 - Opportunity — and increasing necessity — to extract meaning
- Mediation of human interactions by computers
 - Opportunity for the computer in the loop to do much more
- Growing role of language in human-computer interaction

Further motivation for CL

One reason for studying language — and for me personally the most compelling reason — is that it is tempting to regard language, in the traditional phrase, as a “mirror of mind”.

Chomsky, 1975

For the same reason, computational linguistics is a compelling way to study psycholinguistics and language acquisition.

Sometimes, the best way to understand something is to build a model of it.

What I cannot create, I do not understand. Feynman, 1988

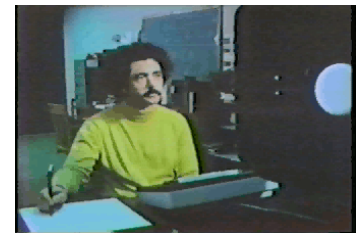
Early history: 50s and 60s

- Foundational work on automata, formal languages, probabilistic modeling, and information theory
- First speech systems (Davis et al., Bell Labs)
- MT heavily funded by military — huge overconfidence
- But using machines dumber than a pocket calculator
- Little understanding of syntax, semantics, pragmatics
- ALPAC report (1966): crap, this is really hard!



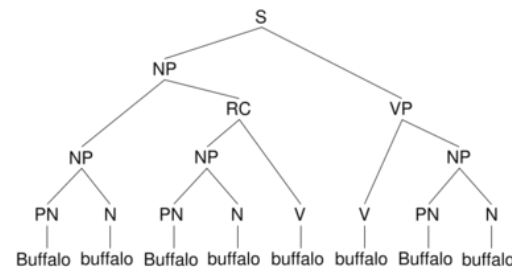
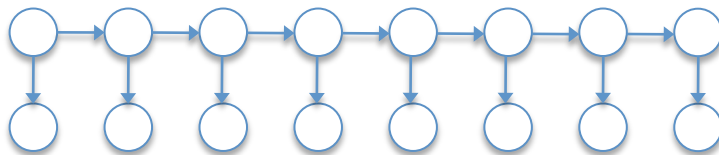
Refocusing: 70s and 80s

- Foundational work on speech recognition: stochastic modeling, hidden Markov models, the “noisy channel”
- Ideas from this work would later revolutionize NLP!
- Logic programming, rules-driven AI, deterministic algorithms for syntactic parsing (e.g., LFG)
- Increasing interest in natural language understanding: SHRDLU, LUNAR, CHAT-80
- But symbolic AI hit the wall: “AI winter”



The statistical revolution: 90s

- Influx of new ideas from EE & ASR: probabilistic modeling, corpus statistics, supervised learning, empirical evaluation
- New sources of data: explosion of machine-readable text; human-annotated training data (e.g., the Penn Treebank)
- Availability of much more powerful machines
- Lowered expectations: forget full semantic understanding, let's do text cat, part-of-speech tagging, NER, and parsing!



The rise of the machines: 00s

- Consolidation of the gains of the statistical revolution
- More sophisticated statistical modeling and machine learning algorithms: MaxEnt, SVMs, Bayes Nets, LDA, etc.
- Big big data: 100x growth of web, massive server farms
- Focus shifting from supervised to *unsupervised* learning
- Revived interest in higher-level semantic applications









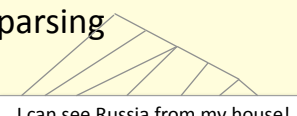







Subfields and tasks

mostly solved

making good progress

still really hard

<p>Spam detection</p> <p>OK, let's meet by the big ... </p> <p>D1ck too small? Buy V1AGRA ... </p>	<p>Sentiment analysis</p> <p>The pho was authentic and yummy. </p> <p>Waiter ignored us for 20 minutes. </p>	<p>Semantic search</p> <p>people protesting globalization <input type="text"/> Search</p> <p> ...demonstrators stormed IMF offices...</p>
<p>Text categorization</p> <p>Phillies shut down Rangers 2-0 SPORTS</p> <p>Jobless rate hits two-year low BUSINESS</p>	<p>Coreference resolution</p> <p></p> <p>Obama told Mubarak he shouldn't run again.</p>	<p>Question answering (QA)</p> <p>Q. What currency is used in China? </p> <p>A. The yuan</p>
<p>Part-of-speech (POS) tagging</p> <p>ADJ ADJ NOUN VERB ADV</p> <p>Colorless green ideas sleep furiously.</p>	<p>Word sense disambiguation (WSD)</p> <p>I need new batteries for my <i>mouse</i>. </p>	<p>Textual inference & paraphrase</p> <p>T. Thirteen soldiers lost their lives ...</p> <p>H. Several troops were killed in the ... YES</p>
<p>Named entity recognition (NER)</p> <p>PERSON ORG LOC</p> <p>Obama met with UAW leaders in Detroit ...</p>	<p>Syntactic parsing</p> <p></p> <p>I can see Russia from my house!</p>	<p>Summarization</p> <p>Sheen continues rant against ...  Sheen is nuts</p>
<p>Information extraction (IE)</p> <p>You're invited to our bunga bunga party, Friday May 27 at 8:30pm in Cordura Hall  Party May 27 add</p>	<p>Machine translation (MT)</p> <p>Our specialty is panda fried rice. </p> <p>我们的专长是熊猫炒饭</p>	<p>Discourse & dialog</p> <p>Where is Thor playing in SF? </p> <p> Metreon at 4:30 and 7:30</p>

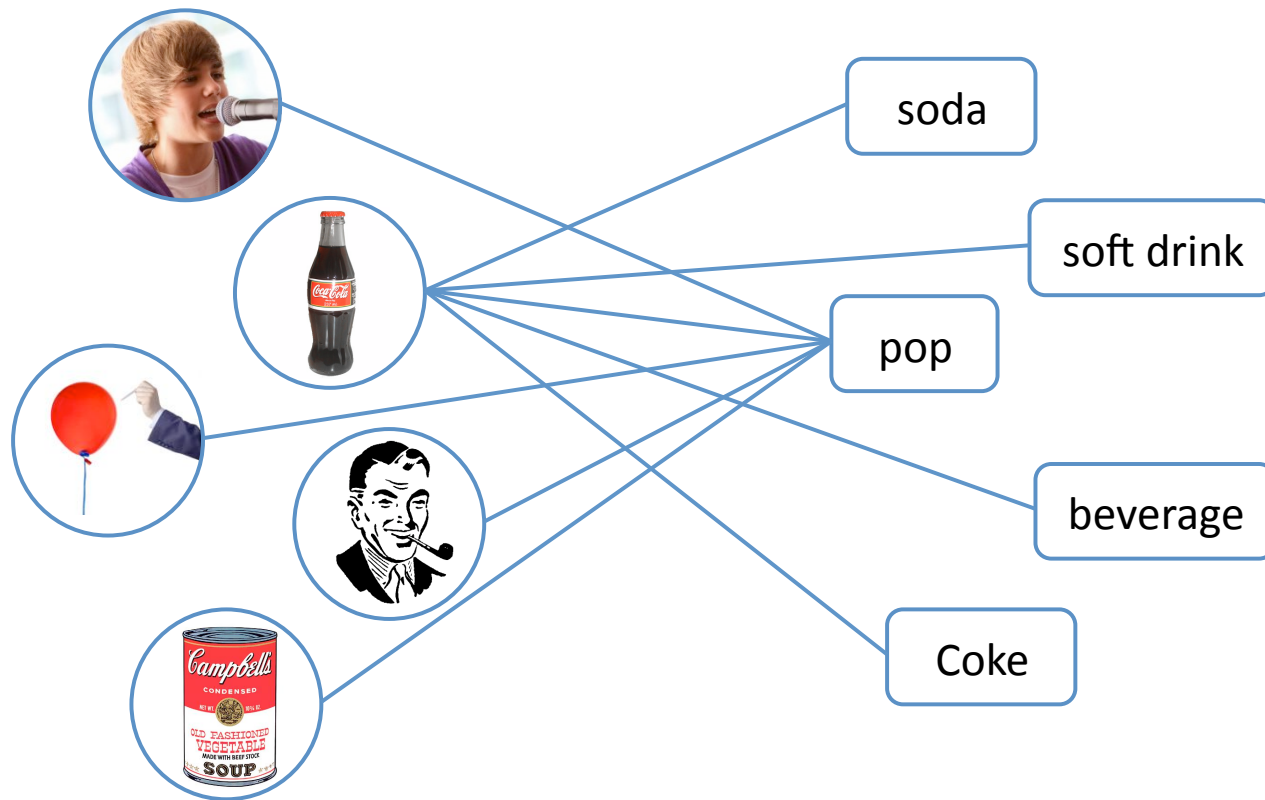
Why is NLP hard?

Natural language is:

- highly ambiguous at all levels
- complex, with recursive structures and coreference
- subtle, exploiting context to convey meaning
- fuzzy and vague
- involves reasoning about the world
- part of a social system: persuading, insulting, amusing, ...

(Nevertheless, simple features often do half the job!)

Meanings and expressions



One meaning, many expressions

To build a shopping search engine, you need to extract product information from vendors' websites:

Image Capture Device: 1.68 million pixel 1/2-inch CCD sensor

Image Capture Device Total Pixels Approx. 3.34 million Effective Pixels Approx. 3.24 million

Image sensor Total Pixels: Approx. 2.11 million-pixel

Imaging sensor Total Pixels: Approx. 2.11 million 1,688 (H) x 1,248 (V)

CCD Total Pixels: Approx. 3,340,000 (2,140[H] x 1,560 [V])

Effective Pixels: Approx. 3,240,000 (2,088 [H] x 1,550 [V])

Recording Pixels: Approx. 3,145,000 (2,048 [H] x 1,536 [V])

These all came from the same vendor's website!

One meaning, many expressions

Or consider a semantic search application:

Russia increasing price of gas for Georgia	Search
Russia hits Georgia with huge rise in its gas bill	
Russia plans to double Georgian gas price	
Russia gas monopoly to double price of gas	
Gazprom confirms two-fold increase in gas price for Georgia	
Russia doubles gas bill to “punish” neighbour Georgia	
Gazprom doubles Georgia's gas bill	

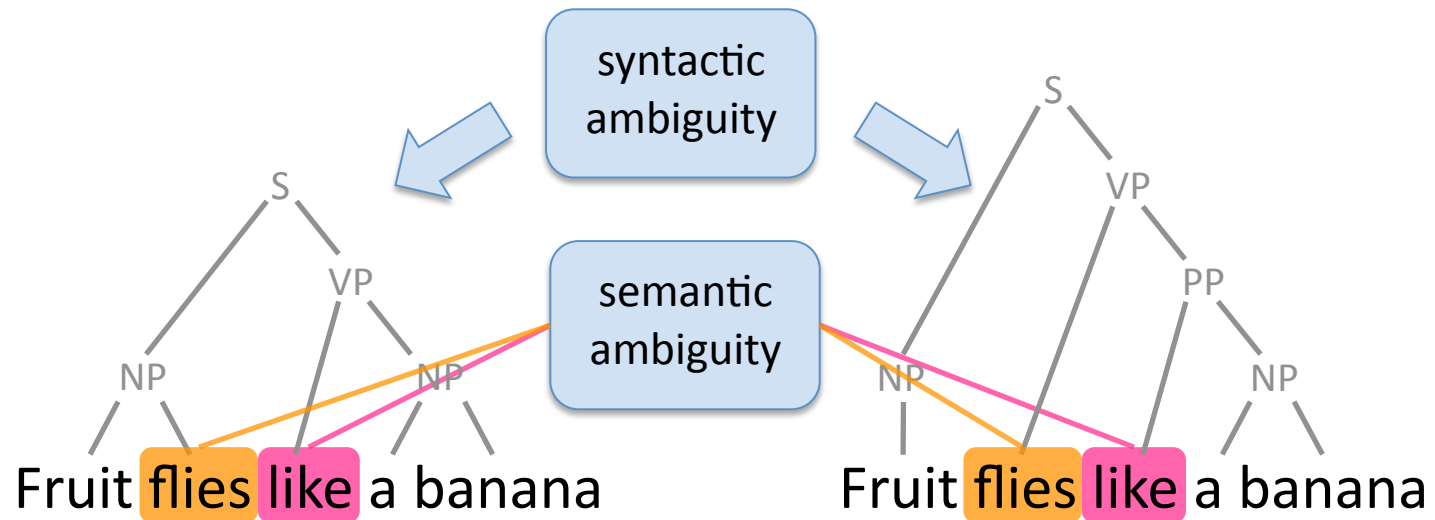
One expression, many meanings



(C) 2003 Ryan North

www.qwantz.com

Syntactic & semantic ambiguity



Ambiguous headlines

Minister Accused Of Having 8 Wives In Jail

May 21, 2007 06:49 AM



ATLANTA (AP) -- A tra
served two years in pri
has been jailed again fo
marry more women.

Bishop Anthony Owens,
Ga., is in a Gwinnett Co
four women claimed he
after being released fro



Teacher Strikes Idle Kids

China to Orbit Human on Oct. 15

Red Tape Holds Up New Bridges

Hospitals Are Sued by 7 Foot Doctors

Juvenile Court to Try Shooting Defendant

Local High School Dropouts Cut in Half

Police: Crack Found in Man's Buttocks

OK, why *else* is NLP hard?

Oh so many reasons!

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

segmentation issues

the New York-New Haven Railroad
the New York-New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
retweet
bromance
teabagger

garden path sentences

The man who hunts ducks out on weekends.
The cotton shirts are made from grows here.

tricky entity names

... a mutation on the *for* gene ...
Where is *A Bug's Life* playing ...
Most of *Let It Be* was recorded ...

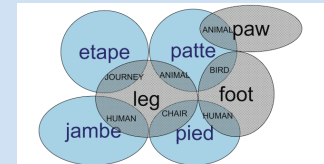
world knowledge

Mary and Sue are sisters.
Mary and Sue are mothers.

prosody

I never said *she* stole my money.
I never said she *stole* my money.
I never said she stole *my* money.

lexical specificity



But that's what makes it fun!

So, how to make progress?

- The task is difficult! What tools do we need?
 - Knowledge about language
 - Knowledge about the world
 - A way to combine knowledge sources
- The answer that's been getting traction:
 - **probabilistic models** built from language data
 - $P(\text{"maison"} \rightarrow \text{"house"})$ **high**
 - $P(\text{"L'avocat général"} \rightarrow \text{"the general avocado"})$ **low**
- Some think this is a fancy new "A.I." idea
 - But really it's an old idea stolen from the electrical engineers ...

Machine translation (MT)

美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击後，关岛经保持高度戒备。



The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

- The classic acid test for natural language processing.
- Requires capabilities in both interpretation and generation.
- About \$10 billion spent annually on human translation.

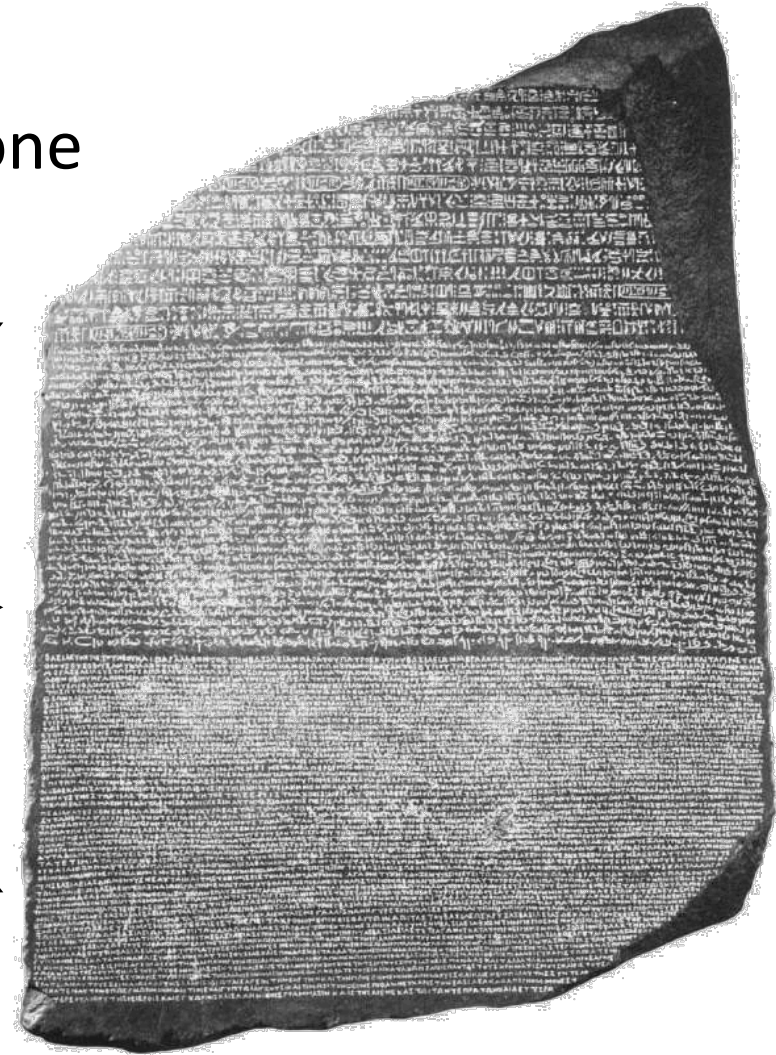
Empirical solution

Parallel Texts: The Rosetta Stone

Hieroglyphs

Demotic

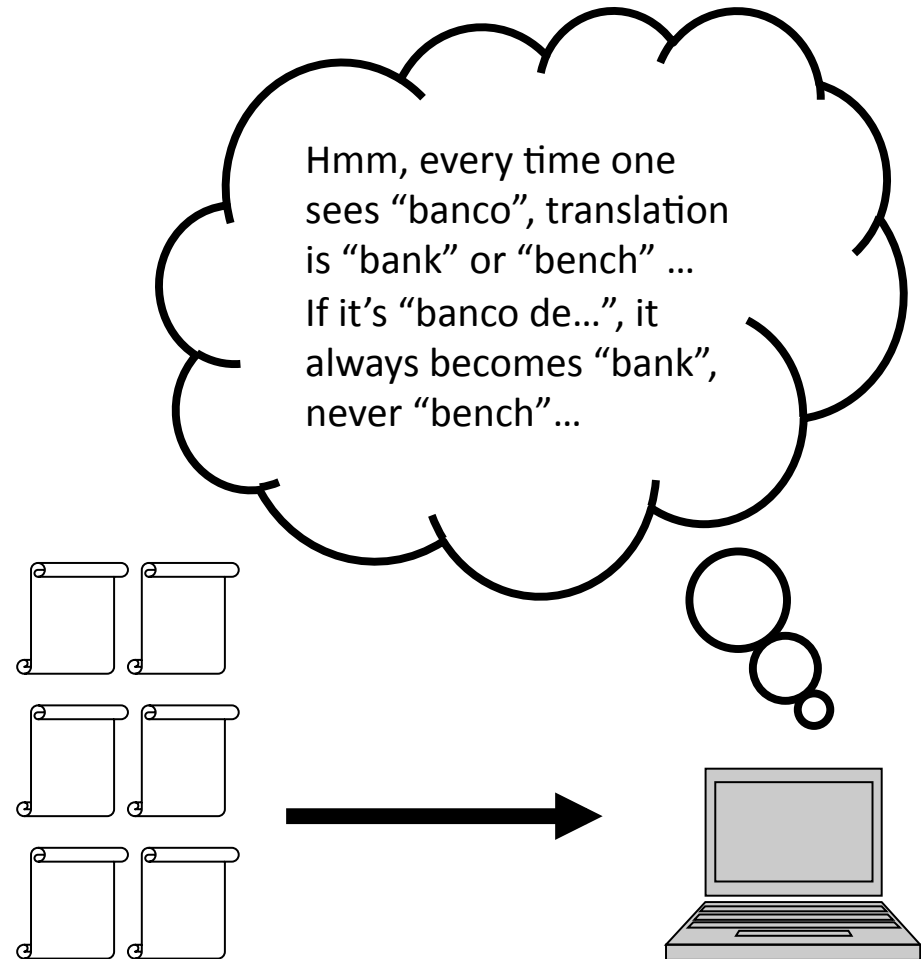
Greek



Empirical solution

Parallel Texts:

- Hong Kong Legislation
- Macao Legislation
- Canadian Parliament Hansards
- United Nations Reports
- European Parliament
- Instruction Manuals
- Multinational company websites



Sindarin-English

I amar prestar aen.

The world is changed.

Han mathon ne nen.

I feel it in the waters.

Han mathon ne chae.

I feel it in the earth.

A han noston ned 'wilith.

I smell it in the air.

Fellowship of the Rings movie script



Statistical MT

Suppose we had a probabilistic model of translation
 $P(e|f)$

Example: suppose f is *de rien*

$P(\textit{you're welcome} | \textit{de rien}) = 0.45$

$P(\textit{nothing} | \textit{de rien}) = 0.13$

$P(\textit{piddling} | \textit{de rien}) = 0.01$

$P(\textit{underpants} | \textit{de rien}) = 0.000000001$

Then the best translation for f is $\operatorname{argmax}_e P(e|f)$

A Bayesian approach

$$\hat{e} = \operatorname{argmax}_e P(e | f)$$

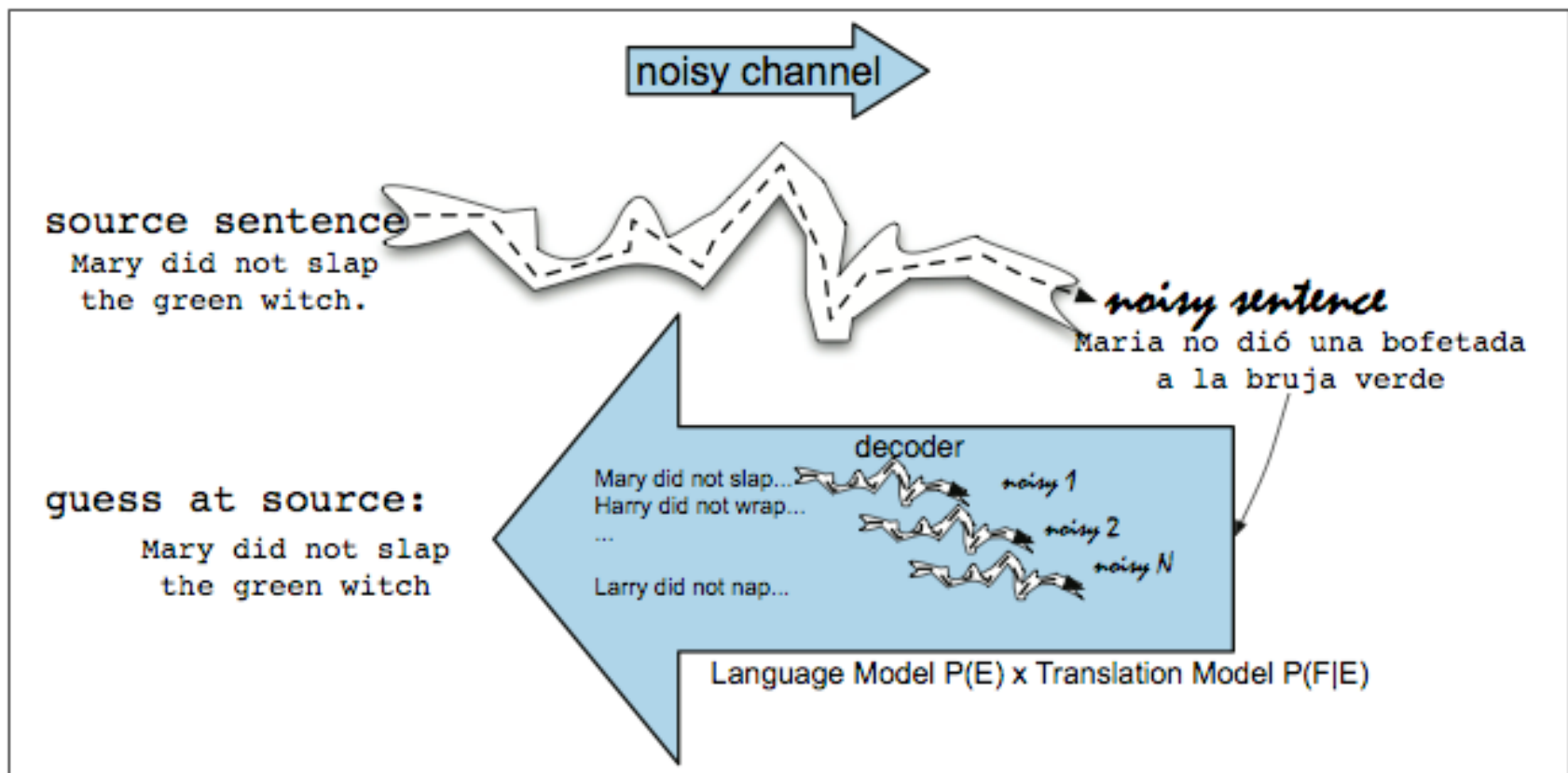
$$= \operatorname{argmax}_e \frac{P(f | e) P(e)}{P(f)}$$

$$= \operatorname{argmax}_e P(f | e) P(e)$$

translation model
(fidelity)

language model
(fluency)

The “noisy channel” model



Language models (LMs)

- Noisy channel model requires language model $P(e)$
- LM tells us which sentences seem likely or “good”
- Given some candidate translations, LM helps with:
 - word choice (“shrank from” or “shrank of”?)
 - word ordering (“tough decisions” or “decisions tough”?)

sentence	$P(e)$
He shrank from tough decisions.	1.89e-11
He shrank from important decisions.	9.46e-12
He shrank of tough decisions.	7.11e-16
He shrank from decisions tough.	3.21e-17

Statistical language models

- Where will the language model come from?
- We'll build it by counting things in corpus data!
- Statistical estimation of model parameters
- But we can't just count whole sentences

sentence	count	P(e)
He shrank from tough decisions.	1/49208	2.03e-05
He shrank from important decisions.	0/49208	0
He shrank of tough decisions.	0/49208	0
He shrank from decisions tough.	0/49208	0

too high!

too low!

N-gram language models

- Instead, we'll break things into pieces

$P(\text{He shrank from tough decisions}) =$

$$P(\text{He} | \bullet) \times P(\text{shrank} | \text{He}) \times P(\text{from} | \text{shrank}) \times \dots \times P(\text{decisions} | \text{tough})$$

- This is called a bigram language model
- We can estimate bigram probabilities from corpus

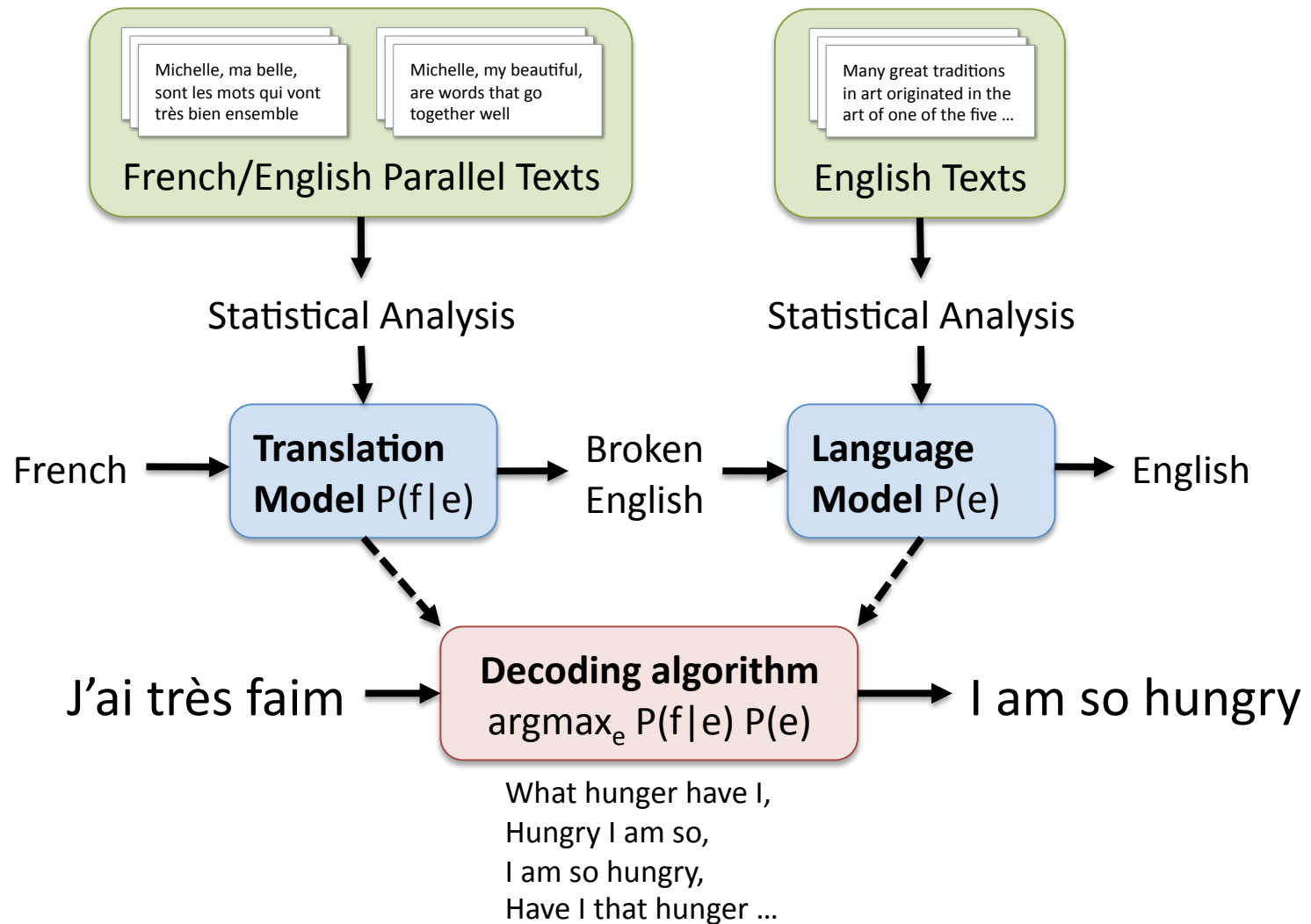
w_1	w_2	$C(w_1)$	$C(w_1 w_2)$	$P(w_2 w_1)$
•	He	49208	978	0.0199
He	shrank	53142	21	0.0004
shrank	from	122	17	0.1393
from	tough	18777	184	0.0098

Statistical translation models

- Noisy channel also needs translation model $P(f|e)$
- Similar strategy: break sentence *pairs* into phrases
- Count co-occurring pairs in a large parallel corpus
- (But I'll skip the gory details ...)

e	f	C(e)	C(e, f)	P(f e)
he shrank	il lui répugnait	17	6	0.3529
from	de	27111	17855	0.6586
from	des	27111	6434	0.2373
tough decisions	décisions difficiles	98	81	0.8265

Statistical MT Systems



Applications of the noisy channel

This model can be applied to many different problems!

$$\hat{e} = \operatorname{argmax}_e P(x|e) P(e)$$

Channel model

speech production

OCR

typing with spelling errors

translating to English

Language model

English words

English words

English words

English words

(Widely used at Google, for example)

If you like NLP / CompLing ...

- learn Java or Python (and play with [JavaNLP](#) or [NLTK](#))
- study logic, probability, statistics, linear algebra
- get some exposure to linguistics (LING1, ...)
- study AI and machine learning (CS121, CS221, CS229)
- read [Jurafsky & Martin](#) or [Manning & Schütze](#)
- CS124: From Language to Information (Jurafsky)
- CS224N: Natural Language Processing (Manning)
- CS224S: Speech Recognition & Synthesis (Jurafsky)
- CS224U: Natural Language Processing (MacCartney)

One more for the road

